Pan Cancer MK2 Survival Analysis

Karthik Suresh

3/11/2020

Updates in Revision 2:

1. We updated the Cox PH model to the model that was used for the LUAD dataset.
2. We updated the figures for the HR across cancer datasets to also include metrics of the Cox PH models used in each dataset.

The MasterMK2data.csv file contains MK2 transcript levels + clinical data across a variety of cancer datasets. Please see the MK2_panCA_annotation pdf for the code that we used to pull the clinical data from the TCGA. The MK2 transcript data was pulled from OncoLnc (manually).

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=70), tidy=TRUE)
```

Create MK2 categories (low vs. high)

An added issue here is that we need to define quantiles WITHIN tumor groups. So the quantile for each tumor group is different.

```
MK2_data <- MK2_data[!is.na(MK2_data$Expression), ]
# create MK2 quantiles
bottom_quant <<- 0.33
top_quant <<- 0.66

BelowBottomQuantile <- function(cancer.type, x) {
    if (x < quantile(MK2_data[MK2_data$cancer.type.x == cancer.type, ]$Expression,
        bottom_quant))
        return(1) else return(0)
}
AboveTopQuantile <- function(cancer.type, x) {
    if (x > quantile(MK2_data[MK2_data$cancer.type.x == cancer.type, ]$Expression,
        top_quant))
        return(1) else return(0)
}
MK2_data$MK2_Expression_bottomQ <- mapply(BelowBottomQuantile, MK2_data$cancer.type.x,
    MK2_data$Expression)
MK2_data$MK2_Expression_topQ <- mapply(AboveTopQuantile, MK2_data$cancer.type.x,
    MK2_data$Expression)


MK2_data_tb <- MK2_data
MK2_data_tb$MK2_Expression_topQ <- as.numeric(MK2_data_tb$MK2_Expression_topQ)
MK2_data_tb$MK2_Expression_bottomQ <- as.numeric(MK2_data_tb$MK2_Expression_bottomQ)
```

```
# re-categorize the variable as MK2_tv where 0 = lower 10th, 1= top
# 10th
MK2_data_tb$MK2_tv <- mapply(function(bottomQ, topQ) if (topQ == 1) return("High") else return("Low"),
    MK2_data_tb$MK2_Expression_bottomQ, MK2_data_tb$MK2_Expression_topQ)
MK2_data_tb$MK2_tv <- as.factor(MK2_data_tb$MK2_tv)

MK2_data_tb <- MK2_data_tb[!is.na(MK2_data_tb$time), ]
# create a 2 year censor
MK2_data_tb$death_at_2year <- mapply(function(dead, time) if (dead == 1 &
    time < 366 * 2) return(1) else if (dead == 0 & time > 366 * 2) return(0) else if (dead ==
    0 & time < 366 * 2) return(0) else if (dead == 1 & time > 366 * 2) return(0),
    MK2_data_tb$censor, as.numeric(MK2_data_tb$time))
MK2_data_tb$death_at_2year <- as.numeric(unlist(as.character(MK2_data_tb$death_at_2year)))
```

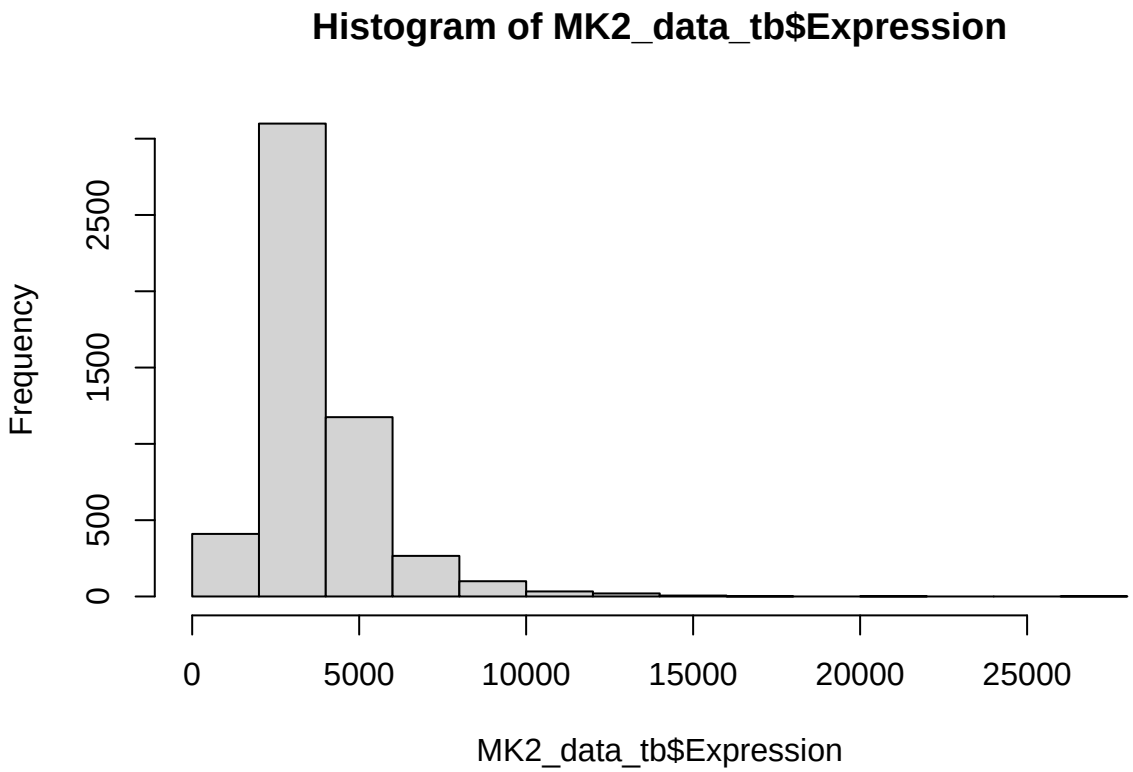
```
## Warning: NAs introduced by coercion
```

```
MK2_data_tb <- MK2_data_tb[MK2_data_tb$death_at_2year == 0 | MK2_data_tb$death_at_2year ==
    1, ]
```

Run Cox PH on the entire dataset

So, here we look at the effect of MK2 on the entire dataset, adjusting for cancer type and stage.

```
hist(MK2_data_tb$Expression)
```




```
MK2_data_tb$Expression_logt <- log(MK2_data_tb$Expression)

coxph_full <- coxph(Surv(time, death_at_2year) ~ Expression_logt + ajcc_pathologic_stage_combined +
    age_at_diagnosis + gender + smoking + as.factor(cancer.type.x), data = MK2_data_tb)
```

```
cox.zph(coxph_full)
```

```
##                                 chisq df       p
## Expression_logt                 0.453  1    0.50
## ajcc_pathologic_stage_combined  0.557  1    0.46
## age_at_diagnosis               17.907  1 2.3e-05
## gender                          1.237  1    0.27
## smoking                         1.343  1    0.25
## as.factor(cancer.type.x)       46.437 12 5.8e-06
## GLOBAL                         67.180 17 6.6e-08
```

```
summary(coxph_full)
```

```
## Call:
## coxph(formula = Surv(time, death_at_2year) ~ Expression_logt +
##     ajcc_pathologic_stage_combined + age_at_diagnosis + gender +
##     smoking + as.factor(cancer.type.x), data = MK2_data_tb)
##
##   n= 5078, number of events= 989
##    (35 observations deleted due to missingness)
##
##                                           coef exp(coef) se(coef)      z
## Expression_logt                        0.17235   1.18810  0.09891  1.742
## ajcc_pathologic_stage_combinedLate Stage  1.03678   2.82011  0.07077 14.650
## age_at_diagnosis                       0.02365   1.02393  0.00299  7.910
## gendermale                             0.05464   1.05616  0.06985  0.782
## smoking1                              -0.08319   0.92018  0.09720 -0.856
## as.factor(cancer.type.x)BRCA          -1.82178   0.16174  0.20620 -8.835
## as.factor(cancer.type.x)COAD          -0.75287   0.47101  0.16045 -4.692
## as.factor(cancer.type.x)ESCA           0.29541   1.34367  0.18692  1.580
## as.factor(cancer.type.x)KIRC          -0.72921   0.48229  0.15386 -4.739
## as.factor(cancer.type.x)KIRP          -1.06740   0.34390  0.23383 -4.565
## as.factor(cancer.type.x)LIHC           0.15580   1.16859  0.15296  1.019
## as.factor(cancer.type.x)LUAD          -0.13029   0.87784  0.14033 -0.928
## as.factor(cancer.type.x)LUSC           0.12708   1.13551  0.13467  0.944
## as.factor(cancer.type.x)PAAD           1.13055   3.09736  0.15347  7.367
## as.factor(cancer.type.x)READ          -1.31750   0.26780  0.29513 -4.464
## as.factor(cancer.type.x)SKCM          -0.64071   0.52692  0.15691 -4.083
## as.factor(cancer.type.x)STAD           0.20768   1.23082  0.13384  1.552
##                                          Pr(>|z|)
## Expression_logt                            0.0814 .
## ajcc_pathologic_stage_combinedLate Stage  < 2e-16 ***
## age_at_diagnosis                       2.57e-15 ***
## gendermale                                 0.4340
## smoking1                                   0.3921
## as.factor(cancer.type.x)BRCA            < 2e-16 ***
## as.factor(cancer.type.x)COAD           2.70e-06 ***
## as.factor(cancer.type.x)ESCA               0.1140
## as.factor(cancer.type.x)KIRC           2.14e-06 ***
## as.factor(cancer.type.x)KIRP           5.00e-06 ***
## as.factor(cancer.type.x)LIHC               0.3084
## as.factor(cancer.type.x)LUAD               0.3532
## as.factor(cancer.type.x)LUSC               0.3454
## as.factor(cancer.type.x)PAAD           1.75e-13 ***
```

```
## as.factor(cancer.type.x)READ             8.04e-06 ***
## as.factor(cancer.type.x)SKCM             4.44e-05 ***
## as.factor(cancer.type.x)STAD               0.1207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                        exp(coef) exp(-coef) lower .95
## Expression_logt                           1.1881     0.8417    0.9787
## ajcc_pathologic_stage_combinedLate Stage  2.8201     0.3546    2.4549
## age_at_diagnosis                          1.0239     0.9766    1.0179
## gendermale                                1.0562     0.9468    0.9210
## smoking1                                  0.9202     1.0867    0.7606
## as.factor(cancer.type.x)BRCA              0.1617     6.1829    0.1080
## as.factor(cancer.type.x)COAD              0.4710     2.1231    0.3439
## as.factor(cancer.type.x)ESCA              1.3437     0.7442    0.9315
## as.factor(cancer.type.x)KIRC              0.4823     2.0734    0.3567
## as.factor(cancer.type.x)KIRP              0.3439     2.9078    0.2175
## as.factor(cancer.type.x)LIHC              1.1686     0.8557    0.8659
## as.factor(cancer.type.x)LUAD              0.8778     1.1392    0.6668
## as.factor(cancer.type.x)LUSC              1.1355     0.8807    0.8721
## as.factor(cancer.type.x)PAAD              3.0974     0.3229    2.2928
## as.factor(cancer.type.x)READ              0.2678     3.7341    0.1502
## as.factor(cancer.type.x)SKCM              0.5269     1.8978    0.3874
## as.factor(cancer.type.x)STAD              1.2308     0.8125    0.9468
##                                        upper .95
## Expression_logt                           1.4423
## ajcc_pathologic_stage_combinedLate Stage  3.2397
## age_at_diagnosis                          1.0299
## gendermale                                1.2111
## smoking1                                  1.1133
## as.factor(cancer.type.x)BRCA              0.2423
## as.factor(cancer.type.x)COAD              0.6451
## as.factor(cancer.type.x)ESCA              1.9382
## as.factor(cancer.type.x)KIRC              0.6520
## as.factor(cancer.type.x)KIRP              0.5438
## as.factor(cancer.type.x)LIHC              1.5771
## as.factor(cancer.type.x)LUAD              1.1558
## as.factor(cancer.type.x)LUSC              1.4785
## as.factor(cancer.type.x)PAAD              4.1843
## as.factor(cancer.type.x)READ              0.4776
## as.factor(cancer.type.x)SKCM              0.7166
## as.factor(cancer.type.x)STAD              1.6000
##
## Concordance= 0.741  (se = 0.007 )
## Likelihood ratio test= 751.5  on 17 df,   p=<2e-16
## Wald test            = 629.1  on 17 df,   p=<2e-16
## Score (logrank) test = 753.2  on 17 df,   p=<2e-16
```

Collect Cox PH info

Now, we run our pre-specified Cox PH model iteratively, for each dataset, and the functions below return either the model result or model metrics.

```r
# Some functions to pull CoxPH models and metrics for all cancers.

returnModel <- function(cancer.type) {
    surv_df <- MK2_data_tb[MK2_data_tb$cancer.type.x == cancer.type, ]
    MK2_model <- coxph(Surv(time, death_at_2year) ~ MK2_Expression_topQ +
        ajcc_pathologic_stage_combined + age_at_diagnosis + gender + smoking,
        data = surv_df)

    return(MK2_model)

}

returnModel_metrics <- function(cancer.type) {
    surv_df <- MK2_data_tb[MK2_data_tb$cancer.type.x == cancer.type, ]
    MK2_model <- coxph(Surv(time, death_at_2year) ~ MK2_Expression_topQ +
        ajcc_pathologic_stage_combined + age_at_diagnosis + gender + smoking,
        data = surv_df)

    return(list(anova(MK2_model), cox.zph(MK2_model)))

}

returnModel_MK2cont <- function(cancer.type) {
    surv_df <- MK2_data_tb[MK2_data_tb$cancer.type.x == cancer.type, ]
    MK2_model <- coxph(Surv(time, death_at_2year) ~ Expression_logt + ajcc_pathologic_stage_combined +
        age_at_diagnosis + gender + smoking, data = surv_df)

    return(MK2_model)

}
returnModel_MK2cont_metrics <- function(cancer.type) {
    surv_df <- MK2_data_tb[MK2_data_tb$cancer.type.x == cancer.type, ]
    MK2_model <- coxph(Surv(time, death_at_2year) ~ Expression_logt + ajcc_pathologic_stage_combined +
        age_at_diagnosis + gender + smoking, data = surv_df)

    return(list(anova(MK2_model), cox.zph(MK2_model)))

}
```

Now that we have some prepared functions, we use them to call the model and retrieve model metrics.

Cox PH Model 1: MK2 as a dichotomous variable + covariates (including stage) Cox PH Model 2: MK2 as a continuous variable + covariates (including stage)

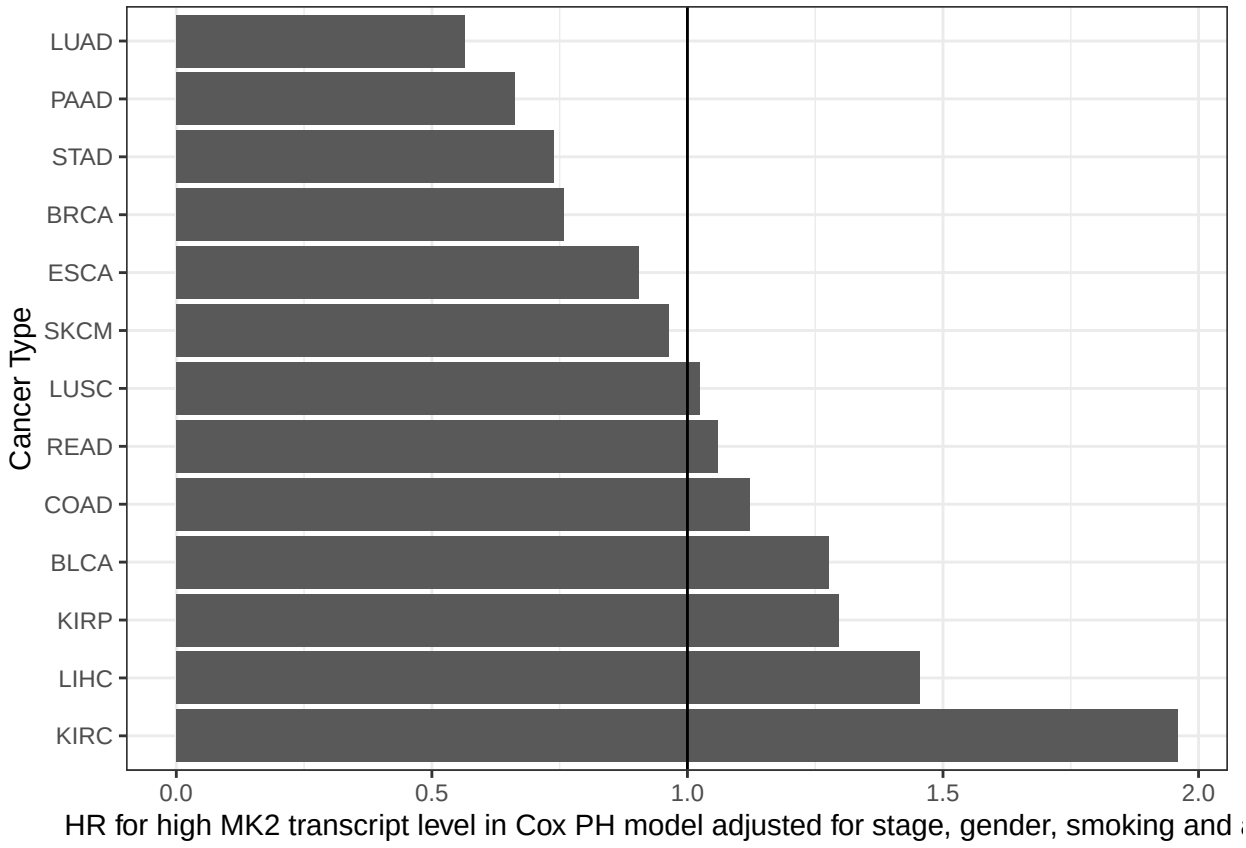
Cancer type is not included because, in essence, we're stratifying by cancer type here.

```r
# return model fit parameters (cox.zph)
MK2_model_df <- MK2_means_wide


# model MK2 as a dichotomous var
MK2_model_df$model1.results <- lapply(MK2_model_df$cancer.type.x, returnModel)


MK2_model_df$model1.metrics <- lapply(MK2_model_df$cancer.type.x, returnModel_metrics)
```

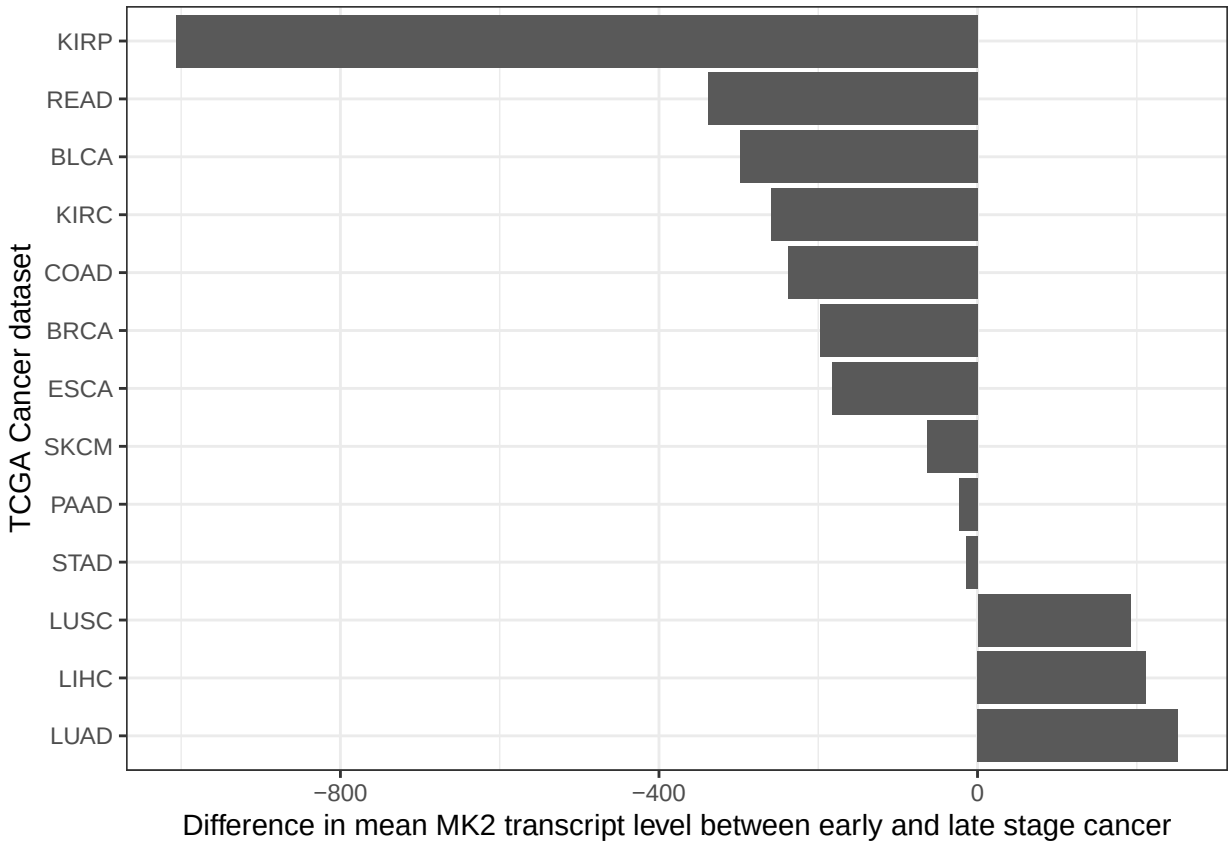
```
MK2_model_df$cox.zph_MK2 <- sapply(MK2_model_df$model1.metrics, function(x) return(x[[2]][[1]][13]))
MK2_model_df$cox.zph_global <- sapply(MK2_model_df$model1.metrics, function(x) return(x[[2]][[1]][18]))
MK2_model_df$wald_pval <- sapply(MK2_model_df$model1.results, function(x) return(broom::glance(x)$p.valu


MK2_model_df$MK2 <- lapply(MK2_model_df$model1.results, function(x) return((x$coefficients[[1]][1])))
MK2_model_df$MK2.lci <- lapply(MK2_model_df$model1.results, function(x) return(exp(confint(x)[1])))
MK2_model_df$MK2.uci <- lapply(MK2_model_df$model1.results, function(x) return(exp(confint(x)[6])))

MK2_model_df$MK2 <- exp(as.numeric(MK2_model_df$MK2))


# model MK2 as a continuous var
MK2_model_df$model2.results <- lapply(MK2_model_df$cancer.type.x, returnModel_MK2cont)
MK2_model_df$MK2_model2 <- lapply(MK2_model_df$model2.results, function(x) return(x$coefficients[[1]][1]
MK2_model_df$MK2_model2.lci <- lapply(MK2_model_df$model2.results, function(x) return(exp(confint(x)[1]
MK2_model_df$MK2_model2.uci <- lapply(MK2_model_df$model2.results, function(x) return(exp(confint(x)[6]
MK2_model_df$MK2_model2 <- exp(as.numeric(MK2_model_df$MK2_model2))

MK2_model_df$model2.metrics <- lapply(MK2_model_df$cancer.type.x, returnModel_MK2cont_metrics)
MK2_model_df$cox.zph_MK2_model2 <- sapply(MK2_model_df$model2.metrics,
    function(x) return(x[[2]][[1]][13]))
MK2_model_df$cox.zph_global_model2 <- sapply(MK2_model_df$model2.metrics,
    function(x) return(x[[2]][[1]][18]))
MK2_model_df$wald_pval_model2 <- sapply(MK2_model_df$model2.results, function(x) return(broom::glance(x
```

Graphical output: graphs of MK2 transcript levels and HR for MK2 across datasets

```
ggplot(MK2_model_df, aes(x = reorder(cancer.type.x, -as.numeric(MK2)),
    y = as.numeric(MK2))) + geom_bar(stat = "identity") + geom_hline(yintercept = 1) +
    theme_bw() + coord_flip() + ylab("HR for high MK2 transcript level in Cox PH model adjusted for stag
    xlab("Cancer Type")
```

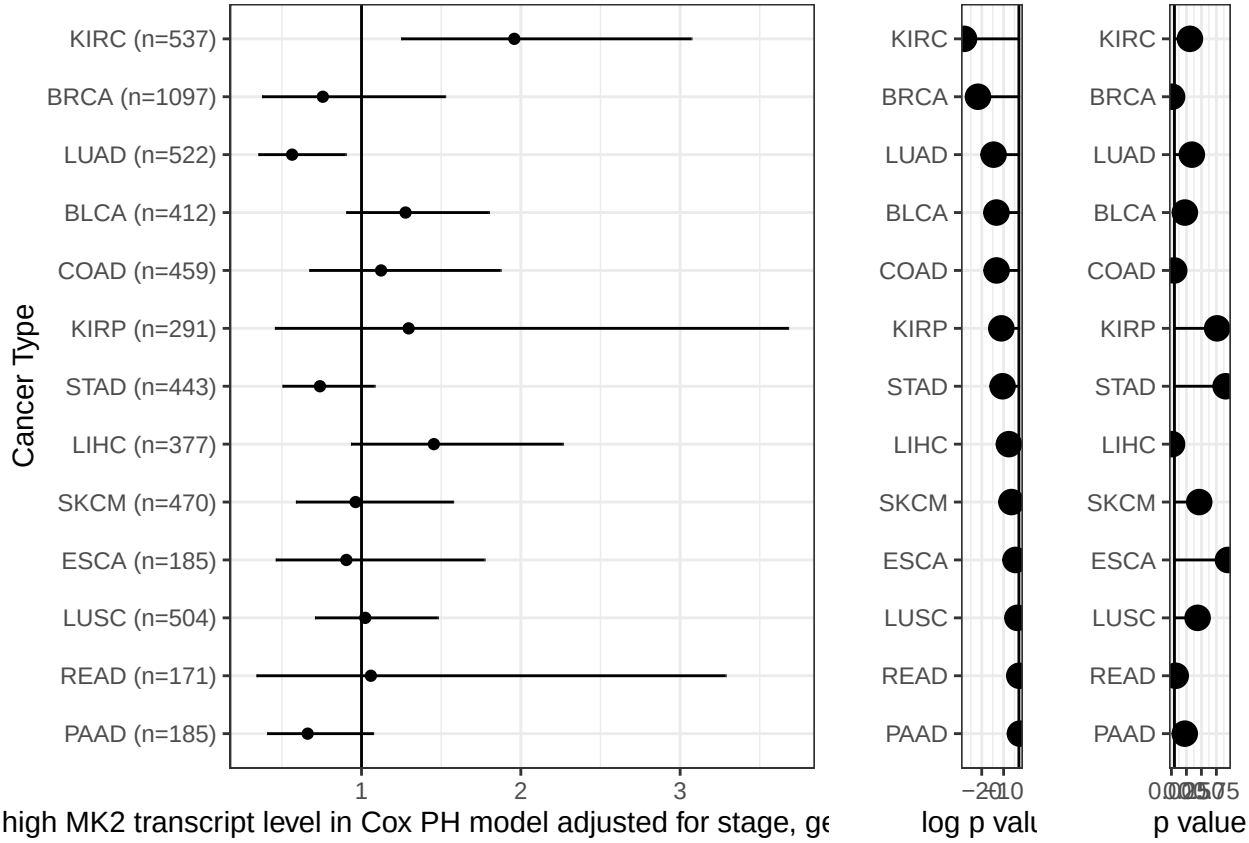



```
ggplot(MK2_model_df, aes(x = reorder(cancer.type.x, -diff), y = diff)) +
    geom_bar(stat = "identity") + theme_bw() + coord_flip() + ylab("Difference in mean MK2 transcript le
    xlab("TCGA Cancer dataset")
```



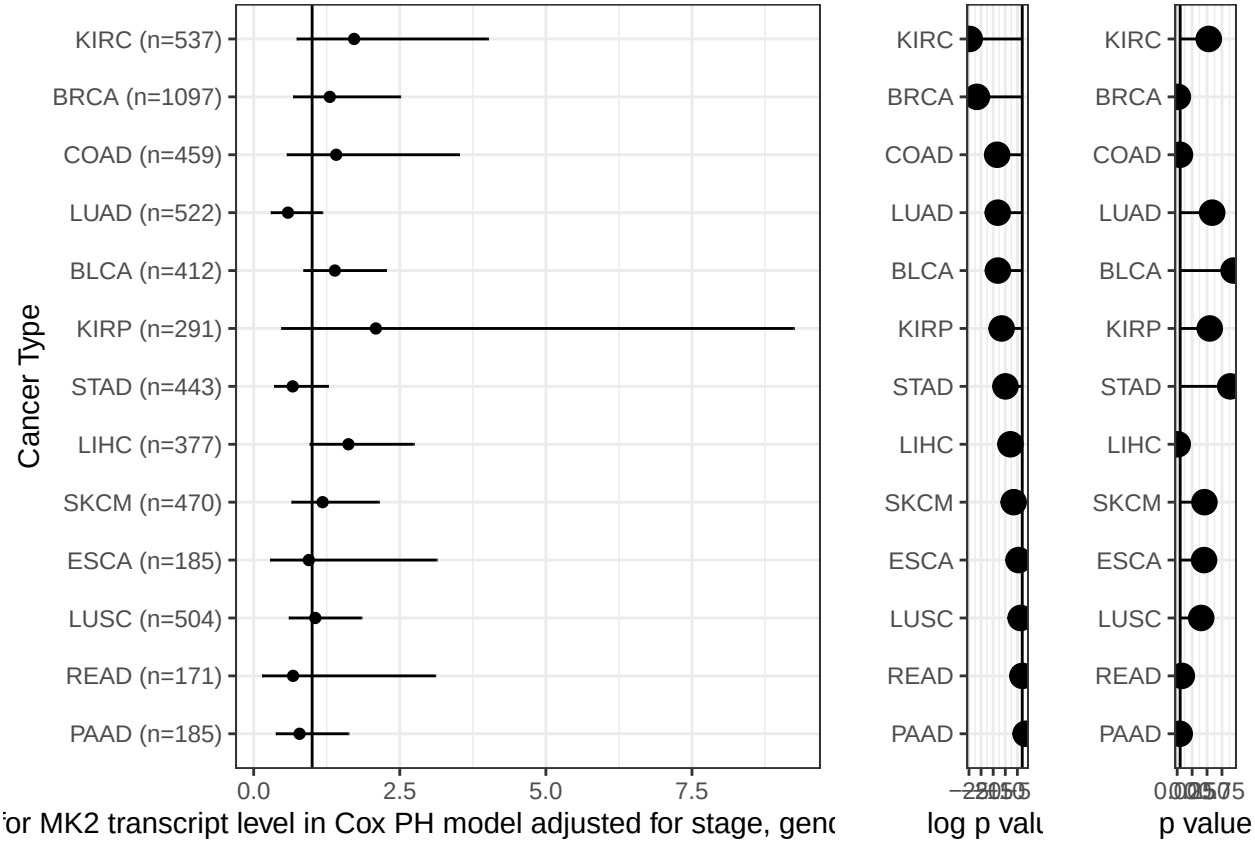

```
MK2_model_df$cancer.type.x.label <- sapply(MK2_model_df$cancer.type.x,
    function(x) return(pts_percatype[pts_percatype$cancer_type == x, ]$num))
MK2_model_df$cancer.type.x.label <- paste0(MK2_model_df$cancer.type.x,
    " (n=", MK2_model_df$cancer.type.x.label, ")")


# Model 1: MK2 as a dichotomous variable

HR_plot <- ggplot(MK2_model_df, aes(x = reorder(cancer.type.x.label, -wald_pval),
    y = as.numeric(MK2))) + geom_point() + geom_errorbar(aes(ymin = as.numeric(MK2.lci),
    ymax = as.numeric(MK2.uci)), width = 0) + geom_hline(yintercept = 1) +
    theme_bw() + ylab("HR for high MK2 transcript level in Cox PH model adjusted for stage, gender, smo
    xlab("Cancer Type") + coord_flip()


model_stats_df <- MK2_model_df[c("cancer.type.x", "cox.zph_MK2", "wald_pval")]
model_stats_df <- arrange(model_stats_df, wald_pval)
metrics_plot_wald <- ggplot(model_stats_df, aes(x = reorder(cancer.type.x,
    -wald_pval), y = log(wald_pval))) + geom_point(size = 4) + geom_hline(yintercept = log(0.05)) +
    geom_segment(aes(x = cancer.type.x, xend = cancer.type.x, y = log(0.05),
        yend = log(wald_pval))) + theme_bw() + xlab("") + ylab("log p value") +
    coord_flip()

metrics_plot_PH <- ggplot(model_stats_df, aes(x = reorder(cancer.type.x,
    -wald_pval), y = cox.zph_MK2)) + geom_point(size = 4) + geom_hline(yintercept = 0.05) +
    geom_segment(aes(x = cancer.type.x, xend = cancer.type.x, y = 0.05,
        yend = cox.zph_MK2)) + theme_bw() + xlab("") + ylab("p value") +
```

```
    coord_flip()


gridExtra::grid.arrange(HR_plot, metrics_plot_wald, metrics_plot_PH, ncol = 3,
    widths = c(4, 1, 1))
```




```
# model 2 results plots

model2_stats_df <- MK2_model_df[c("cancer.type.x", "cox.zph_MK2_model2",
    "wald_pval_model2")]

HR_plot_model2 <- ggplot(MK2_model_df, aes(x = reorder(cancer.type.x.label,
    -wald_pval_model2), y = as.numeric(MK2_model2))) + geom_point() + geom_errorbar(aes(ymin = as.numer
    ymax = as.numeric(MK2_model2.uci)), width = 0) + geom_hline(yintercept = 1) +
    theme_bw() + ylab("HR for MK2 transcript level in Cox PH model adjusted for stage, gender, smoking
    xlab("Cancer Type") + coord_flip()

metrics_plot_wald_model2 <- ggplot(model2_stats_df, aes(x = reorder(cancer.type.x,
    -wald_pval_model2), y = log(wald_pval_model2))) + geom_point(size = 4) +
    geom_hline(yintercept = log(0.05)) + geom_segment(aes(x = cancer.type.x,
    xend = cancer.type.x, y = log(0.05), yend = log(wald_pval_model2))) +
    theme_bw() + xlab("") + ylab("log p value") + coord_flip()

metrics_plot_PH_model2 <- ggplot(model2_stats_df, aes(x = reorder(cancer.type.x,
    -wald_pval_model2), y = cox.zph_MK2_model2)) + geom_point(size = 4) +
    geom_hline(yintercept = 0.05) + geom_segment(aes(x = cancer.type.x,
    xend = cancer.type.x, y = 0.05, yend = cox.zph_MK2_model2)) + theme_bw() +
```

```
    xlab("") + ylab("p value") + coord_flip()

gridExtra::grid.arrange(HR_plot_model2, metrics_plot_wald_model2, metrics_plot_PH_model2,
    ncol = 3, widths = c(4, 1, 1))
```



R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

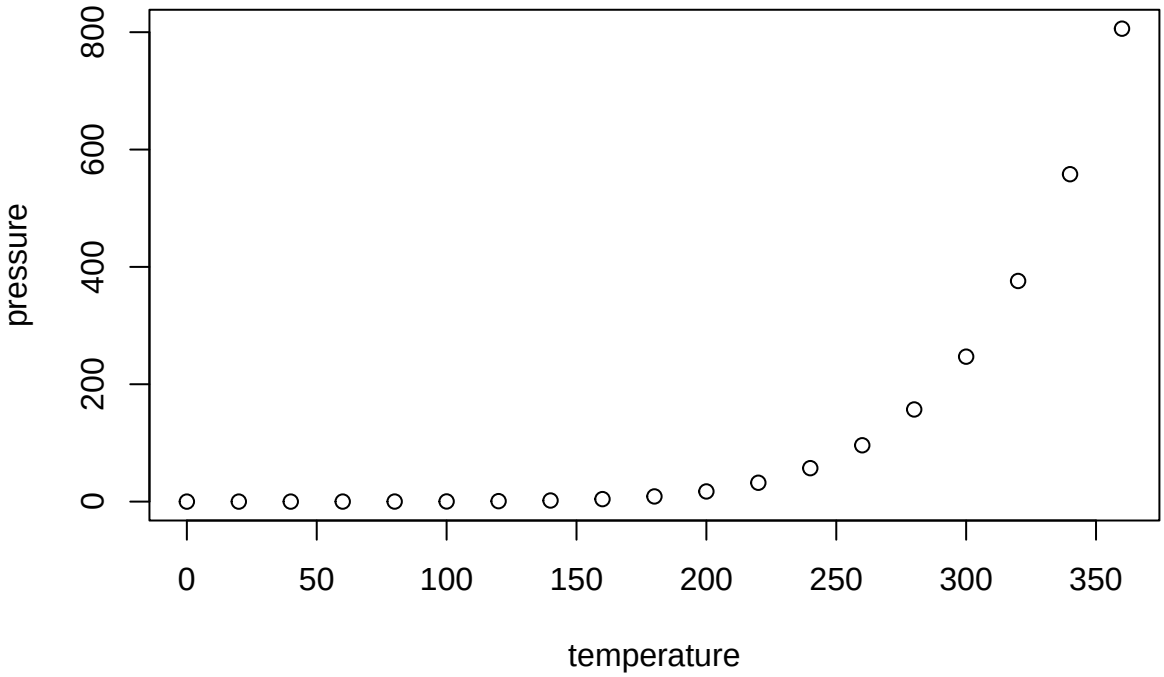
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.