Name: Sanjana Yasna

The Following is on an .ipynb file, but the r-code I use is pasted in with relevant outputs

3.1 (p. 121 ISLR)

Sales has no relationship with newspaper sales (null hypothesis accepted), but sales has a relationship with TV and radio advertising (null hypothesis rejected).

We see that the t-statistic is relatively large in comparison to the standard error for both TV and radio to sales, meaning the slope and intercept of the regression line is likely not close to 0. This also explains the very low p-values (well below even a 1% cutoff) that indicate there is a very low chance that this relationship between the predictor and response in these two cases is due to chance. The odd one out is sales to newspaper, which has a very high p-value and a t statistic that isn't as many magnitudes larger than the standard error as the relationship with TV and sales, so it fits criteria for keeping hte null hypothesis.

3.4 (p. 122 ISLR)

(a)

RSS would be lower for cubic regression than linear regression

Cubic regression aims to reduce the error term ϵ and since the curve of fit has three fitted coefficients as opposed to the 1 for linear regression, it typically gets lower variance in its fit and better minimizes standard error of its reference points. So, RSS will eventually be lower in cubic since the model has more opportunity to overfit based on subtleties.

(b)

Test set RSS would be lower in linear regression than cubic regression

The underlying relationship in the data is linear, so RSS training on cubic regression has greater chance of overfit that would lead to poor test performance than linear regression

(c)

Again, cubic regression would yield lower RSS than linear regression

Similar reasoning to (a). if the data is very non-linear and follows a cubic or spline pattern, RSS may be much lower for cubic regression in comparison to linear regression

(d)

Unknown, as is the underlying distribution is unknown, test performance can't be said

Cubic regression may generally do better than linear regression, but if the test set is still mostly linear, linear regression would do better.

3.6 (p. 123 ISLR)

Linear regression line prediction at given point is: $\hat{y}_i = \hat{B}_0 + \hat{B}_0 x_i$

$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$

We want to minimize the error $\epsilon_i = y_i - \hat{y}_i$, so aim for $y_i = \hat{y}_i$ to try for zero error and set the regression line equation to y_i instead:

$y_i = \hat{B}_0 + \hat{B}_1 x_i$ And after substituting for B_0:

$y_i = \bar{y} - \hat{B}_1 \bar{x} + \hat{B}_1 x_i$

If we allow $x_i = \bar{x}$ to cancel terms out we are left with just $y_i = \bar{y}$

So if $x_i = \bar{x}$ and $y_i = \bar{y}$, the prediction $\hat{y}_i$ will equal the true value y_i at a certain point, minimizing RSS.

$(\bar{x}, \bar{y})$ also makes sense in direct context of just minimizing the individual $\hat{B}_1$ and $\hat{B}_0$ terms:

$\hat{B}_1 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$

So when $x_i = \bar{x}$ and $y_i = \bar{y}$, $\hat{B}_1 = 0$ at that point, and $\hat{B}_0 = \bar{y}$, which would push the prediction $\hat{y}_i = \bar{y}$

3.8 (p. 123-124 ISLR)

(a)

There is a clear negative relationship between horsepower and mpg that doesn't seem to be by chance, as the p-value is a mere 2.2e-16 and the F-statistic is 599.7, a strong indication we should reject the null hypothesis. Horsepower of 98 is associated with around a predicted 24.47 mpg, and 95% confidence interval is [23.97308, 24.96108] and 95% prediction interval is [14.8094, 34.12476]. There appears to be a somewhat strong relationship between the response and predictor since the R^2 statistic is 0.6049, meaning almost 2/3rds of the variance in horsepower is due to the linear regression on mpg. The RSS is 4,906, and given the mean mpg is 23.44, there is around a 21% error in prediction terms on average, indicative of a decent fit.

While it is hard to say exactly what F-statistic cutoff we should use as a must for a strong rejection of null hypothesis especially since this is a small dataset of only 392 observations, the F-statistic is far greater than 1 so we can assume a significant relationship at this point. The prediction interval is significantly wider than the confidence interval (which usually is the case). The slope of the fitted line is around -0.15 (negative relationship) and the intercept is around 39.94.

Here's the R code I used and summary statistics:

SUMMARY STATISTICS:

```
{r}
require(ISLR)
Auto = ISLR::Auto
lm_fit <- lm(mpg ~ horsepower, data = Auto)
summary(lm_fit)
```

OUTPUT: Call: lm(formula = mpg ~ horsepower, data = Auto)

Residuals: Min 1Q Median 3Q Max -13.5710 -3.2592 -0.3435 2.7630 16.9240

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861 0.717499 55.66 <2e-16 *** horsepower -0.157845 0.006446 -24.49 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049 F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

CONFIDENCE AND PRED INTERVALS:

```
{r}
to_predict = data.frame(horsepower = c(98))
#confidence and prediction by default does 95% intervals, so I assume I don't need
to specify interval bounds here?
predict(lm_fit, newdata =  to_predict, interval = "confidence")
predict(lm_fit, newdata =  to_predict, interval = "prediction")
```

OUTPUT: fit lwr upr 1 24.46708 23.97308 24.96108 fit lwr upr 1 24.46708 14.8094 34.12476

(b)

```
{r}
plot(horsepower, mpg)
abline(lm_fit)
```


(c)

Linear regression may not be the most appropriate method of fitting the data, but it isn't unreasonable. The third figure (Scale-Location) has a conical pattern of standarzied residual error terms as you fit higher horsepower values to mpg, and this lack of heteroscadicity violates an assumption of linear regression, so the model doesn't perfectly suit the conditions of ground truth data. The second graph (Q-Q Residuals) have the desired linear pattern between theoretical quantiles and standardized residuals and that means that horsepower and mpg follow very similar distributions. Despite this similarity between input and target variables, the first plot (Residuals vs Fitted) has only a slightly parabolic shape in the residual terms the model tries to fit; while it doesn't have the undesired horizontal residual fit line that centers at 0, it makes it clear that the data doesn't follow a strict linear relationship (how strict is hard to tell).In the fourth plot, all data points seem to be within the Cook's distance influence range and so the model's fit wasn't heavily skewed by any outliers (which is good for regression).

Diagnostic plots below:

3.10 (p. 124-125 ISLR)

(a)

CODE (R):

```
{r}
multi_lr <- lm(Sales ~ Price + Urban + US, data = carseats)
summary(multi_lr)
```

SUMMARY STATS: Call: lm(formula = Sales ~ Price + Urban + US, data = carseats)

Residuals: Min 1Q Median 3Q Max -6.9206 -1.6220 -0.0564 1.5786 7.0581

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469 0.651012 20.036 < 2e-16 ***

Price -0.054459 0.005242 -10.389 < 2e-16 ***

UrbanYes -0.021916 0.271650 -0.081 0.936 ***

USYes 1.200573 0.259042 4.635 4.86e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335 F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

(c)

The fitted equation can be stated as : sales = 13.043469 - 0.054459Price - 0.021916UrbanYes + 1.200573*USYes

(b)

Price coefficient: very small negative relationship to sales (small coefficient), but relationship is signficant since p-value is very small and t-value is rather

UrbanYes Coefficient: very very small negative coefficient, but there appears to likely be no significant relationship to sales since p-value is rather high. (and t-statistic is small and matches distribution expected of null hypothesis) . Assume null hypothesis

UsYes Coefficient: strong positive relationship (low p-value, t-value is high enough to indicate relationship significance)

r-squared is just 0.2335, which is poor overall fit. Residual standard error is 2.472, which is significant given this is in thousands of units (so off by around 2.5k units on average...a 34% error...)

(d)

As discussed in part b above, price and USYes likely have a relationship to sales so I reject null for those two

(e)

CODE

```
{r}
two_lr <- lm(Sales ~ Price + US, data = carseats)
summary(two_lr)
```

OUT Call: lm(formula = Sales ~ Price + US, data = carseats)

Residuals: Min 1Q Median 3Q Max -6.9269 -1.6286 -0.0574 1.5766 7.0515

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079 0.63098 20.652 < 2e-16 ***

Price -0.05448 0.00523 -10.416 < 2e-16 ***

USYes 1.19964 0.25846 4.641 4.71e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354 F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

R-squared is almost identical even with the fit only by the variables we thought significant

(f)

There is no notable improvement in fitting sales on only to USYes and Price. At best, the r-squared value went up by a few hundredths, and the residual standard error went down too by a few hundredths, so there was a very minor improvement on fitting to only price and USYes. Otherwise, both models are mediocre.

(g)

95% confidence intervals for price are [-0.06475984, -0.04419543] and USYes are [0.69151957, 1.70776632]

(h)

Again, we see that in the Residuals vs Leverage plot (fourth figure), while a handful of points get close to Cook's distance, all of the standardized residuals have no significant leverage and saty within Cook's distance. The Residuals vs Fitted curve has residuals constrained to within 5 of the target, and the distribution of residuals seems to be relatively and randomly uniform around 0 (which is an undesired pattern for linear regression) and standardized residual distances fall within 1.5. (Scale-Location graph) There appears to be no significant outliers with much higher residuals than the rest of the data.

DIAGNOSTIC PLOTS BELOW

In [ ]:
```
!pip install -U notebook-as-pdf
```

In [3]:
```
!pip install 'PyPDF2<3.0'
```

```
DEPRECATION: Loading egg at /Users/sanjanayasna/.pyenv/versions/3.12.0/lib/python3.12/site-packages/pyrosetta-2024.18+release.117e2f6f54-py3.12-macosx-14.1-x86_64.egg is deprecated. pip 24.3 will enforce this behaviour change. A possible replacement is to use pip for package installation.. Discussion can be found at https://github.com/pypa/pip/issues/12330
DEPRECATION: Loading egg at /Users/sanjanayasna/.pyenv/versions/3.12.0/lib/python3.12/site-packages/pymsim-4.3.2-py3.12-macosx-14.1-x86_64.egg is deprecated. pip 24.3 will enforce this behaviour change. A possible replacement is to use pip for package installation.. Discussion can be found at https://github.com/pypa/pip/issues/12330
Collecting PyPDF2<3.0
  Downloading pypdf2-2.12.1-py3-none-any.whl.metadata (6.6 kB)
Downloading pypdf2-2.12.1-py3-none-any.whl (222 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 222.8/222.8 kB 3.5 MB/s eta 0:00:00a 0:00:01
Installing collected packages: PyPDF2
  Attempting uninstall: PyPDF2
    Found existing installation: PyPDF2 3.0.1
    Uninstalling PyPDF2-3.0.1:
      Successfully uninstalled PyPDF2-3.0.1
Successfully installed PyPDF2-2.12.1

[notice] A new release of pip is available: 24.0 -> 25.0.1
[notice] To update, run: pip install --upgrade pip
```

In [4]:
```
!jupyter-nbconvert --to PDFviaHTML /Users/sanjanayasna/csc_math_classes/csc293/SCO1/ICO1_Sanjana_Ya
```