This week we're going to practice testing predictive accuracy of a model. See the end of the Correlation notebook for an example!

Primary Exercise

Create an 80/20 train-test split, and use some of the Penguin models from Regressions to predict penguin body mass. Train both on non-standardized data, to make prediction easier.

Measure the RMSE or MSE of these predictions.

Which is more accurate? The flipper model, or the multivariate flipper/species/sex model?

Is this difference statistically significant? You can't directly test the RMSEs, but the RMSEs are computed from the squared error for individual test samples - can you test for significant differences in squared error? What is the appropriate significance test here?

Upload your notebook and PDF to Canvas.

More Fun If you have time - let's try to see bias/variance.s

- Create an empty list. Do the following 1000 times:
- Create an 80/20 train/test split
- Train the model on train
- Predict the ratings on test
- Put the predictions into a frame with columns for iteration, penguin number (row from original data frame; this is the index of your predictions), and body mass
- Append this frame to the list

Use pd.concat to concatenate the list into a frame. Now, if you group by predictor variables, you'll get the average classification for each set of predictors (covariates). You can use this to compute the variance of your model's predictions.

In [46]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_rel
```

In [47]:
```
penguins = pd.read_csv('penguins.csv')
penguins.head()
```

Out[47]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN	2007
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007

In [48]:
```
test = penguins.sample(frac=0.2)
train_mask = pd.Series(True, index=penguins.index)
train_mask[test.index] = False
train = penguins[train_mask]
```

Flipper length and body mass

In [49]:
```
bm_mod = smf.ols('body_mass_g ~ flipper_length_mm', data=train)
bmf = bm_mod.fit()
bmf.summary()
```

Out[49]:

OLS Regression Results

Dep. Variable:	body_mass_g	**R-squared:**	0.762
Model:	OLS	**Adj. R-squared:**	0.761
Method:	Least Squares	**F-statistic:**	866.7
Date:	Thu, 21 Oct 2021	**Prob (F-statistic):**	2.09e-86
Time:	10:04:42	**Log-Likelihood:**	-2018.6
No. Observations:	273	**AIC:**	4041.
Df Residuals:	271	**BIC:**	4048.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5809.2542	341.876	-16.992	0.000	-6482.324	-5136.184
flipper_length_mm	49.8972	1.695	29.440	0.000	46.560	53.234

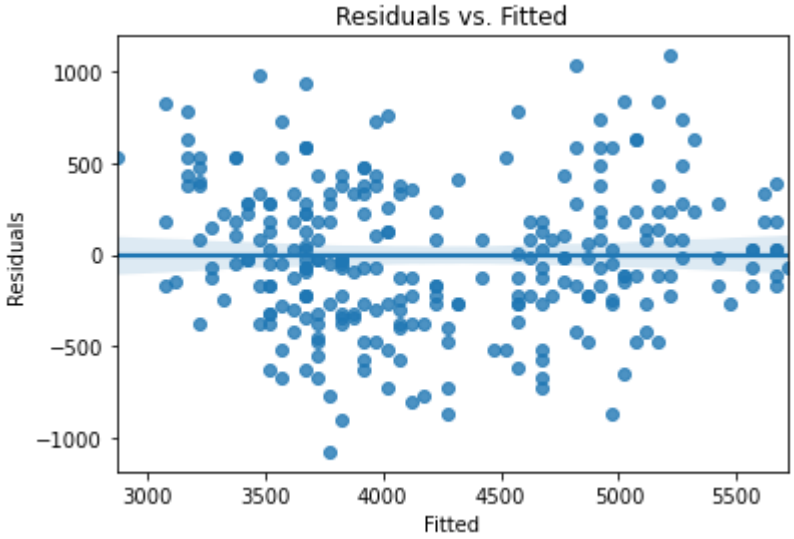
Omnibus:	1.150	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.563	**Jarque-Bera (JB):**	1.243
Skew:	0.146	**Prob(JB):**	0.537
Kurtosis:	2.844	**Cond. No.**	2.89e+03

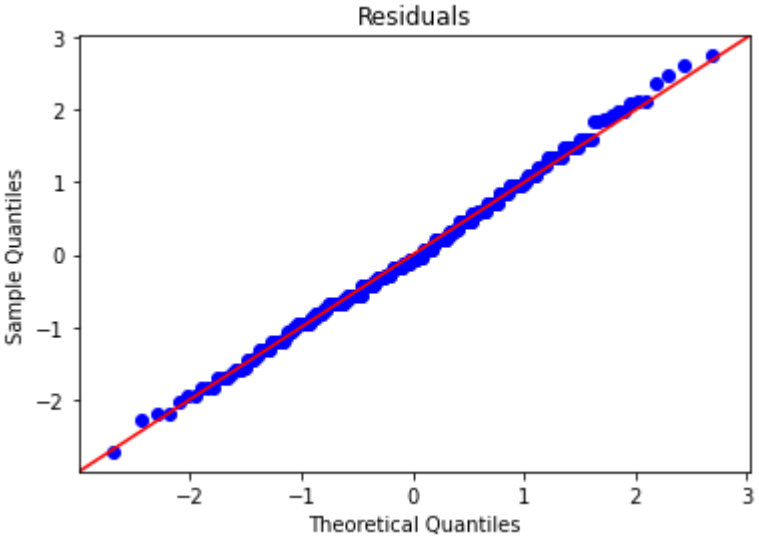
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.89e+03. This might indicate that there are strong multicollinearity or other numerical problems.

In [50]:
```
def plot_lm_diag(fit):
    "Plot linear fit diagnostics"
    sns.regplot(x=fit.fittedvalues, y=fit.resid)
    plt.xlabel('Fitted')
    plt.ylabel('Residuals')
    plt.title('Residuals vs. Fitted')
    plt.show()

    sm.qqplot(fit.resid, fit=True, line='45')
    plt.title('Residuals')
    plt.show()

plot_lm_diag(bmf)
```





In [51]:
```
preds = bmf.predict(test)
test['pred_body_mass_g'] = preds
test
```

Out[51]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	yea
62	Adelie	Biscoe	37.6	17.0	185.0	3600.0	female	200

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	yea
298	Chinstrap	Dream	43.2	16.6	187.0	2900.0	female	200
28	Adelie	Biscoe	37.9	18.6	172.0	3150.0	female	200
117	Adelie	Torgersen	37.3	20.5	199.0	3775.0	male	200
30	Adelie	Dream	39.5	16.7	178.0	3250.0	female	200
...	...	...	...	...	...	...	...	.
278	Chinstrap	Dream	51.3	19.2	193.0	3650.0	male	200
310	Chinstrap	Dream	49.7	18.6	195.0	3600.0	male	200
137	Adelie	Dream	40.2	20.1	200.0	3975.0	male	200
148	Adelie	Dream	36.0	17.8	195.0	3450.0	female	200
140	Adelie	Dream	40.2	17.1	193.0	3400.0	female	200

69 rows × 9 columns

In [52]:
```
test['error_simple'] = test['body_mass_g'] - test['pred_body_mass_g']
np.mean(np.square(test['error_simple']))
```

Out[52]: 154445.33375956767

In [53]:
```
np.sqrt(np.mean(np.square(test['error_simple'])))
```

Out[53]: 392.9953355442882

In [54]:
```
test['error_simple'] = np.square(test['error_simple'])
```

Flipper length, body mass, species and sex

In [55]:
```
bm_mod = smf.ols('body_mass_g ~ flipper_length_mm + species * sex + flipper_length_mm:s
bmf = bm_mod.fit()
bmf.summary()
```

Out[55]:

OLS Regression Results

Dep. Variable:	body_mass_g	**R-squared:**	0.880
Model:	OLS	**Adj. R-squared:**	0.876
Method:	Least Squares	**F-statistic:**	233.5
Date:	Thu, 21 Oct 2021	**Prob (F-statistic):**	1.12e-112
Time:	10:04:43	**Log-Likelihood:**	-1862.9
No. Observations:	264	**AIC:**	3744.
Df Residuals:	255	**BIC:**	3776.

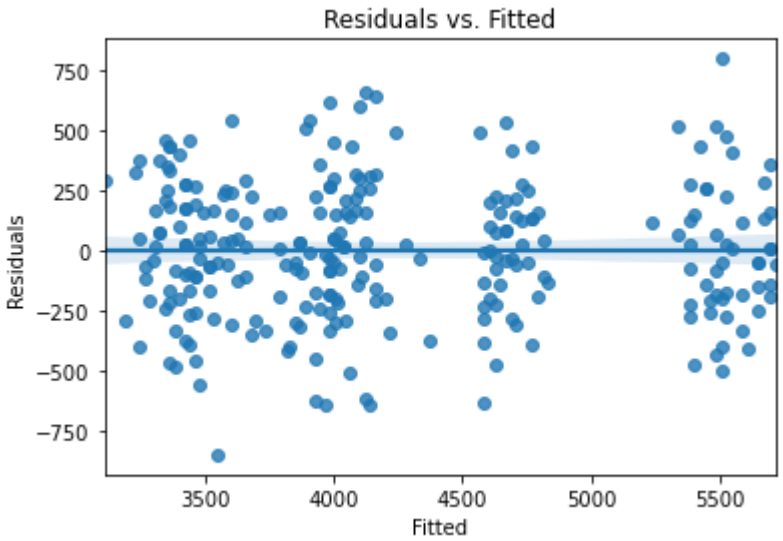
Df Model:	8
Covariance Type:	nonrobust

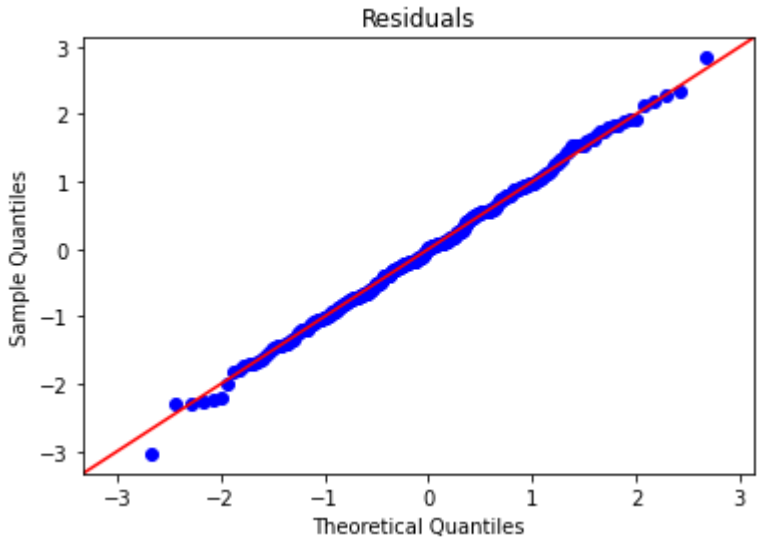
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-281.7953	834.149	-0.338	0.736	-1924.493	1360.902
species[T.Chinstrap]	344.6211	1544.019	0.223	0.824	-2696.031	3385.274
species[T.Gentoo]	538.0591	1549.303	0.347	0.729	-2513.001	3589.119
sex[T.male]	561.0455	57.150	9.817	0.000	448.500	673.591
species[T.Chinstrap]:sex[T.male]	-277.0641	111.452	-2.486	0.014	-496.547	-57.581
species[T.Gentoo]:sex[T.male]	87.7898	99.085	0.886	0.376	-107.340	282.920
flipper_length_mm	19.5044	4.430	4.403	0.000	10.780	28.229
flipper_length_mm:species[T.Chinstrap]	-1.3321	8.086	-0.165	0.869	-17.256	14.592
flipper_length_mm:species[T.Gentoo]	1.3007	7.570	0.172	0.864	-13.606	16.208

Omnibus:	0.081	**Durbin-Watson:**	2.013
Prob(Omnibus):	0.960	**Jarque-Bera (JB):**	0.202
Skew:	0.002	**Prob(JB):**	0.904
Kurtosis:	2.865	**Cond. No.**	2.38e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.38e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

In [56]:
```
plot_lm_diag(bmf)
```





In [57]:
```
preds = bmf.predict(test)
test['pred_complex_model_body_mass_g'] = preds
test
```

Out[57]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	yea
62	Adelie	Biscoe	37.6	17.0	185.0	3600.0	female	200
298	Chinstrap	Dream	43.2	16.6	187.0	2900.0	female	200
28	Adelie	Biscoe	37.9	18.6	172.0	3150.0	female	200
117	Adelie	Torgersen	37.3	20.5	199.0	3775.0	male	200
30	Adelie	Dream	39.5	16.7	178.0	3250.0	female	200
...	...	...	...	...	...	...	...	.
278	Chinstrap	Dream	51.3	19.2	193.0	3650.0	male	200
310	Chinstrap	Dream	49.7	18.6	195.0	3600.0	male	200
137	Adelie	Dream	40.2	20.1	200.0	3975.0	male	200
148	Adelie	Dream	36.0	17.8	195.0	3450.0	female	200
140	Adelie	Dream	40.2	17.1	193.0	3400.0	female	200

69 rows × 11 columns

In [58]:
```
test['error_complex'] = test['body_mass_g'] - test['pred_complex_model_body_mass_g']
np.mean(np.square(test['error_complex']))
```

Out[58]: 90920.15140614418

In [59]:
```
np.sqrt(np.mean(np.square(test['error_complex'])))
```

Out[59]: 301.52968577926816

In [61]:
```
test['error_complex'] = np.square(test['error_complex'])
```

In [63]:
```
test
```

Out[63]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	yea
62	Adelie	Biscoe	37.6	17.0	185.0	3600.0	female	200
298	Chinstrap	Dream	43.2	16.6	187.0	2900.0	female	200
28	Adelie	Biscoe	37.9	18.6	172.0	3150.0	female	200
117	Adelie	Torgersen	37.3	20.5	199.0	3775.0	male	200
30	Adelie	Dream	39.5	16.7	178.0	3250.0	female	200
...	...	...	...	...	...	...	...	.
278	Chinstrap	Dream	51.3	19.2	193.0	3650.0	male	200
310	Chinstrap	Dream	49.7	18.6	195.0	3600.0	male	200
137	Adelie	Dream	40.2	20.1	200.0	3975.0	male	200
148	Adelie	Dream	36.0	17.8	195.0	3450.0	female	200
140	Adelie	Dream	40.2	17.1	193.0	3400.0	female	200

69 rows × 12 columns

In [62]:
```
ttest_rel(test["error_complex"], test["error_simple"])
```

Out[62]:
```
Ttest_relResult(statistic=3.3444335274360357, pvalue=0.001344585622003043)
```

Q. Which is more accurate? The flipper model, or the multivariate flipper/species/sex model?

A. The multivariate model has lower RSME and is more accurate.

Q. Is this difference statistically significant? You can't directly test the RMSEs, but the RMSEs are computed from the squared error for individual test samples - can you test for significant differences in squared error? What is the appropriate significance test here?

A. The RSMEs were different. Yes, we can test for significant difference in squared error using Paired -tests.